CLAIMS

1. A method for browser-enhanced web crawling associated with a network of hub processing units coupled to a plurality of information processing units over a network, the method executed by a web crawler on a hub processing unit associated with the network comprising the steps of:

retrieving a document at an address;

loading secondary documents;

sending to one or more information processing units a browser side script to gather metadata; and performing the sub-steps of:

producing a final HTML markup;

analyzing and summarizing the final HTML markup to produce metadata.

2. The method as defined in claim 1, wherein the retrieving a document at an address step further comprises retrieving a document at an address selected from the group of addresses consisting of a nodal address, a network address, a URL and equivalents.

3. The method as defined in claim 1, wherein the analyzing and summarizing step further comprises analyzing and summarizing the whole and complete document.

4. The method as defined in claim 1, further comprising the step of analyzing any image data present in the document and any image date present in the documents utilizing optical character recognition techniques.

5. The method as defined in claim 1, wherein the step of loading secondary documents further comprises the loading of secondary documents including documents selected from the group of documents consisting of in-line frames, frames, images, image maps, applets, audio, video or equivalents.

6. The method as defined in claim 4, wherein the step of analyzing any image data present in the document and any image date present in the documents utilizing optical character recognition techniques further comprises analyzing any images and image maps in the image data to produce text data.

7. The method as defined in claim 1, wherein the retrieving step further comprises performing the sub-steps of:

initializing a first list with seed values;

checking if there are any URLs to be processed and if there are, performing the secondary sub-steps of:

determining if a URL is in a second list; and if it is not in the second list; then performing the tertiary sub-steps of:

inserting the URL into the first list;

scheduling the URL for crawling;

crawling the URL when scheduled to do so;

removing the URL from the first list after the scheduled crawling;

entering the URL into the second list; and

repeating the checking step until there are no more URLs to be processed;

where if the determining step determines that the URL is in the second list then repeating the checking step until there are no more URLs to be processed.

8. The method as defined in claim 7, wherein the sub-step of initializing a first list with seed values further includes the list being a URL pool.

9. The method as defined in claim 7, wherein the sub-step of determining if a URL is in a second list further includes the second list being a visited pool.

10. The method as defined in claim 7, wherein the tertiary sub-step of crawling further comprises the sub-steps of:

issuing an HTTP command to a web server named in the URL;

receiving contents of an HTML page as a result of the issued HTTP command; and

passing on the contents of the HTML page to a Page Rendering subroutine.

11. The method as defined in claim 10, further including the sub-steps performed by the Page Rendering subroutine comprising:

receiving the contents of the HTML page in the Page Rendering subroutine;

building an in-memory representation of a Layout for the HTML page and if more data is needed to properly form the representation, then performing the sub-steps of:

requesting additional web-based information;

gathering this additional web-based information;

inserting any URLs associated with this additional web-based information into the second list and a URL cache;

building a final amended representation; and

forwarding the final amended representation to an Extraction subroutine;

wherein, if no more data is needed to properly form the in-memory representation, then forwarding the in-memory representation to the Extraction subroutine.

12. The method as defined in claim 11, further including the sub-steps performed by the Page Extraction subroutine comprising:

accessing a set of memory structures of the Page Renderer;

copying a text portion of the structures into a text map;

inspecting any in-line GIF and JPEG image references in the memory structures;

extracting alternate text attributes;

adding the alternate text attributes to a text map;

invoking an optical character recognition engine;

analyzing any in-line GIF and JPEG images using the optical character recognition engine for text content;

extracting text content from the GIF and JPEG images;

adding text content from the images to the text map; and

forwarding the text map to a Page Summarizer subroutine.

13. The method as defined in claim 12, further including the sub-steps performed by the Page Summarizer subroutine comprising:

receiving a text map from the Page Extractor subroutine;

processing the text map in an application-specific manner;

applying data extraction patterns to the text map;

translating resultant data from the applying step;

forwarding any URLs present in the text map to a manager subroutine; and

forwarding any extracted data and metadata to application logic.

14. A computer readable medium including programming instructions, the programming instructions including instructions for browser-enhanced web crawling associated with a network of hub processing units coupled to a plurality of information processing units over a network, the browser enhanced web crawling instructions on the computer readable medium comprising:

retrieving instructions for retrieving a document at an address;

loading instructions for loading secondary documents;

sending instructions for sending to one or more information processing units a browser side script to gather metadata;

producing instructions for producing a final HTML markup;

analyzing and summarizing instructions for analyzing and summarizing the final HTML markup to produce the final metadata.

15. The computer readable medium as defined in claim 14, wherein the retrieving instructions for retrieving a document at an address further comprises retrieving instructions for retrieving a document at an address selected from the group of addresses consisting of a nodal address, a network address, a URL and equivalents.

16. The computer readable medium as defined in claim 14, wherein the analyzing and summarizing instructions further comprise analyzing and summarizing instructions for analyzing and summarizing the whole and complete document.

17. The computer readable medium as defined in claim 14, further comprising image analyzing instructions for analyzing any image data present in the document and any image date present in the documents utilizing optical character recognition techniques.

18. The computer readable medium as defined in claim 14, wherein the loading instructions for loading secondary documents further comprises loading instructions for loading of secondary documents including documents selected from the group of documents consisting of in-line frames, frames, images, image maps, applets, audio, video or equivalents.

19. The computer readable medium as defined in claim 17, wherein the analyzing instructions for analyzing any image data present in the document and any image date present in the documents utilizing optical character recognition techniques further comprises analyzing instructions for analyzing any images and image maps in the image data to produce text data.

20. A browser-enhanced web crawling unit associated with a network of a plurality of hub processing units coupled to a plurality of information processing units over a network, the browser enhanced web crawling unit on a hub processing unit comprising:

a retrieval unit for retrieving a document at an address;

a loader for loading secondary documents as required;

an output for sending to one or more information processing units a browser side script to gather metadata;

a producer for producing a final HTML markup; and

a summarizer for analyzing and summarizing the final HTML markup to produce the final metadata.

known to one of ordinary skill in the art, arranged to perform the functions described and the method steps described. The operations of such a computer, as described above, may be according to a computer program contained on a medium for use in the operation or control of the computer, as would be known to one of ordinary skill in the art. The computer medium which may be used to hold or contain the computer program product, may be a fixture of the computer such as an embedded memory or may be on a transportable medium such as a disk, as would be known to one of ordinary skill in the art.

The invention is not limited to any particular computer program or logic or language, or instruction but may be practiced with any such suitable program, logic or language, or instructions as would be known to one of ordinary skill in the art. Without limiting the principles of the disclosed invention any such computing system can include, inter alia, at least a computer readable medium allowing a computer to read data, instructions, messages or message packets, and other computer readable information from the computer readable medium. The computer readable medium may include non-volatile memory, such as ROM, Flash memory, floppy disk, Disk drive memory, CD-ROM, and other permanent storage. Additionally, a computer readable medium may include, for example, volatile storage such as RAM, buffers, cache memory, and network circuits.

Furthermore, the computer readable medium may include computer readable information in a transitory state medium such as a network link and/or a network interface, including a wired network or a wireless network, that allow a computer to read such computer readable information.

What is claimed is: